Mimicking the Thinking Process for Emotion Recognition in Conversation with Prompts and Paraphrasing

Ting Zhang[1] , **Zhuang Chen**[2] , **Ming Zhong**[1*] and **Tieyun Qian**[1,3*]
[1]School of Computer Science, Wuhan University
[2]The CoAI group, DCST, Tsinghua University
[3]Intellectual Computing Laboratory for Cultural Heritage, Wuhan University
tingzhang_17@whu.edu.cn, zhchen-nlp@mail.tsinghua.edu.cn, {clock, qty}@whu.edu.cn

Abstract

Emotion recognition in conversation, which aims to predict the emotion for all utterances, has attracted considerable research attention in recent years. It is a challenging task since the recognition of the emotion in one utterance involves many complex factors, such as the conversational context, the speaker's background, and the subtle difference between emotion labels. In this paper, we propose a novel framework which mimics the thinking process when modeling these factors. Specifically, we first *comprehend the conversational context* with a history-oriented prompt to selectively gather information from predecessors of the target utterance. We then *model the speaker's background* with an experience-oriented prompt to retrieve the similar utterances from all conversations. We finally *differentiate the subtle label semantics* with a paraphrasing mechanism to elicit the intrinsic label related knowledge. We conducted extensive experiments on three benchmarks. The empirical results demonstrate the superiority of our proposed framework over the state-of-the-art baselines.

1 Introduction

Emotion Recognition in Conversation (ERC) is a crucial task for recognizing mental health problems. As reported by the WHO, 1 in every 8 people in the world lives with a mental health problem such as emotional regulation [1]. With the goal of recognizing human emotions, ERC can help discover the negative emotions of speakers and identify potential individuals who may be experiencing mental health issues. The ERC task is also a fundamental step towards human-like artificial intelligence (AI) [Poria *et al.*, 2019b], and has played an important role in many areas that are beneficial to humans such as legal trials [Poria *et al.*, 2019b], empathetic dialog systems [Majumder *et al.*, 2020], health care systems [Pujol *et al.*, 2019], and intelligent assistants [König *et al.*, 2016].

Different from conventional emotion recognition tasks, the emotion of a target utterance in ERC is not self-contained, which indicates that we cannot predict the emotion merely by understanding the utterance itself. Instead, some supplementary information, such as the conversational context, and the speaker's background, is required to accurately identify the emotion conveyed by the utterance. Moreover, the difference between emotion labels like '*sadness*' and '*frustrated*' is often subtle and needs to be carefully distinguished.

Current research direction is mainly towards the modeling of the conversational context without taking the speaker's background into account. Various sequence-based models [Hazarika *et al.*, 2018; Majumder *et al.*, 2019; Hu *et al.*, 2021] and graph-based models [Ghosal *et al.*, 2019; Shen *et al.*, 2021a; Bao *et al.*, 2022; Li *et al.*, 2022; Shen *et al.*, 2021b] built upon pre-trained language models (PLMs) are developed to model contextual interactions between utterances. There is also a growing trend in employing external commonsense knowledge [Sap *et al.*, 2019] to enrich utterance representations [Ghosal *et al.*, 2020; Zhu *et al.*, 2021] or facilitate emotion transition over the conversation graphs [Li *et al.*, 2021; Zhao *et al.*, 2022]. Simply fusing features through network structure falls short of exploiting the knowledge capacity [Liu *et al.*, 2023; Brown *et al.*, 2020; Liu *et al.*, 2021] of the PLMs, and thus a more recent method CISPER [Yi *et al.*, 2022] leverages the prompt-learning paradigm for this purpose. However, CISPER uses the same prompt for all utterances in a dialogue without considering their relations to the target utterance and its speaker.

In this paper, we propose a novel conversational emotion recognition framework which mimics the thinking process of a human being. To understand the emotion conveyed by the target utterance, human beings typically go through the following questions step by step.

1) *What does the speaker say?* People need first to locate and then read the target utterance to understand the utterance itself.
2) *What is the influence of the conversational context on the speaker?* The conversational context may exert a strong influence on the speaker, so it is necessary to obtain relevant information from the dialogue history.
3) *What is the speaker's background?* People need to learn the speaker's background since the speaker often draws

* Corresponding authors.

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

experience in similar situations to express an attitude to a particular utterance.

4) *How does the speaker feel?* People need to differentiate the semantics of the emotion labels for a precise emotion understanding.

To realize the above thinking process, we present a **m**ulti-**p**rompt and **l**abel **p**araphrasing (MPLP) model for emotion recognition in conversation. Our model consists of two stages. At the first stage, the model is trained to understand 1) *What does the speaker say?*. An utterance along with its surrounding context are fed into a PLM for the utterance encoding. In the second stage, the model is further trained to better identify the emotions with a thorough comprehension of 2) 3) 4). While 2) is the focus of all previous methods yet the existing prompt based approach does not handle it well, 3) and 4) are largely neglected by current studies. Our effort is devoted to addressing these issues.

To be specific, to perceive 2) *What is the influence of the conversational context on the speaker?*, we encode the speaker-related information and the history-influenced emotion into **a history-oriented prompt** to comprehend the conversational context. To model 3) *What is the speaker's background?*, we retrieve similar utterances seen in the training set and convert these utterances into **an experience-oriented prompt** to capture the speaker's task-specific experience. To have a deep understanding of 4) *How does the speaker feel?*, we design an auxiliary generation task with the help of **label paraphrasing** from SentiWordNet [Baccianella *et al.*, 2010] to distinguish the subtle semantics of different emotion labels.

In summary, the contributions of this work are threefold. Firstly, we point out the problem of inadequate coverage of the human thinking process in existing methods. Secondly, we propose a multi-prompt and label paraphrasing model to mimic this process in a comprehensive way. Lastly, We demonstrate the effectiveness and the working mechanism of our proposed model via extensive experiments on three commonly-used datasets [2].

2 Related Work

2.1 Emotion Recognition in Conversation

Most existing approaches design sequence-based or graph-based models to tackle the problem of context modeling. ICON [Hazarika *et al.*, 2018] uses GRUs to model the self- and inter-speaker emotional effects. DialogueRNN [Majumder *et al.*, 2019] keeps track of the individual party states by several GRU models. DialogueCRN [Hu *et al.*, 2021] uses LSTM modules to retrieve and integrate contextual emotional clues iteratively. DialogueGCN [Ghosal *et al.*, 2019] models speakers' dependency by applying graph neural networks to a neighbor graph. DAG-ERC [Shen *et al.*, 2021b] and SGED [Bao *et al.*, 2022] treat the conversation as an acyclic directed graph. External knowledge is also widely used in ERC tasks. COSMIC [Ghosal *et al.*, 2020] introduces commonsense knowledge during the sequence modeling procedure. TODKAT [Zhu *et al.*, 2021] combines topic information to reduce the noise of commonsense. SKAIG [Li *et al.*, 2021] and CauAIN [Zhao *et al.*, 2022] classify commonsense elements into different types to enhance emotion transition between utterances. CISPER [Yi *et al.*, 2022] encodes the conversation context and commonsense into prompts. In addition to the model designing, some work [Yang *et al.*, 2022; Song *et al.*, 2022; Li *et al.*, 2022] adopts contrastive or curriculum learning strategies to get better results.

Overall, existing methods suffer from the issue of inadequate coverage of the human thinking process, and we realize this by mimicking the complete process with prompts and paraphrasing.

2.2 Prompt and Paraphrasing

Prompt-based learning is an emerging paradigm in natural language processing. To bridge the gap between the pre-training and fine-tuning, prompt-based methods modify the inputs by appending additional token sequences, which are helpful to elicit knowledge from PLMs [Brown *et al.*, 2020; Lester *et al.*, 2021]. Early models like GPT-3 [Brown *et al.*, 2020] use handcrafted task instructions and demonstrations to construct hard prompts. Recently, there has been a growing trend to explore the potential of continuous prompts [Li and Liang, 2021; Lester *et al.*, 2021]. Paraphrasing is another new learning paradigm to transfer knowledge of a PTM by paraphrasing the key elements and generating a target sentence for the input sentence [Mueller *et al.*, 2022; Zhang *et al.*, 2021].

We introduce a history-oriented prompt, an experience-oriented prompt, and a label paraphrasing into our model, which can better leverage the power of PLMs for the downstream ERC task.

3 Methodology

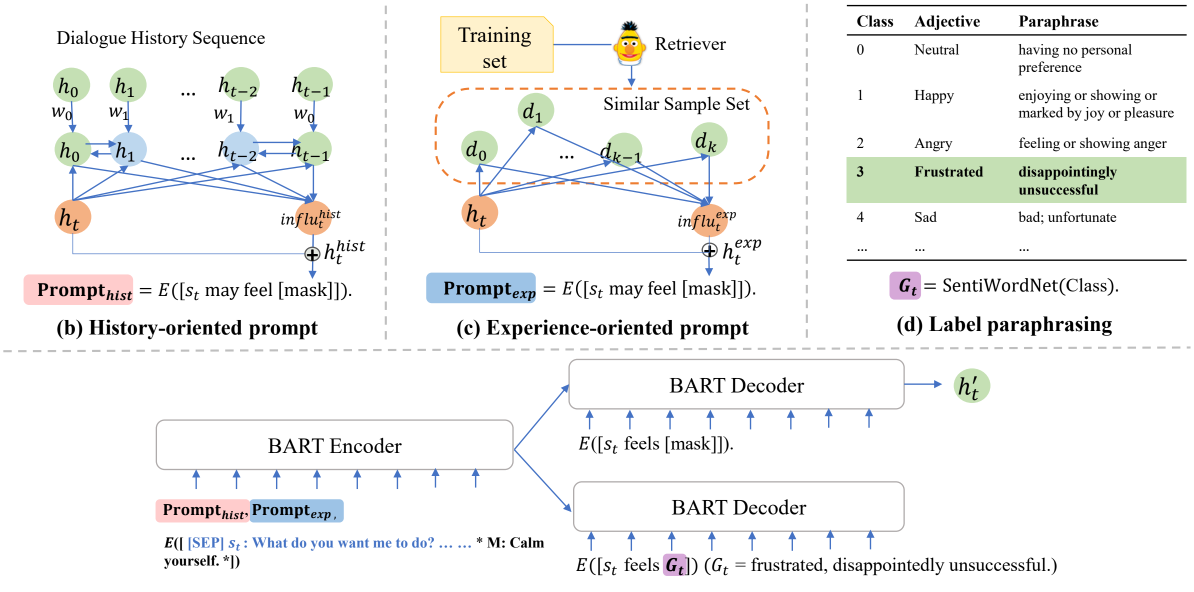
3.1 Problem Definition

In ERC, a conversation is defined as a list of utterances $u_1, u_2, ..., u_N$, where N is the number of utterances. Each utterance u_i consists of n_i tokens, namely $u_i = w_{i1}, w_{i2}, ..., w_{in_i}$. A discrete value $y_i \in Y$ is used to denote the emotion label of u_i, where Y is the set of emotion labels. Each utterance u_i is associated with a speaker $s(u_i)$. The objective of this task is to output the emotion label y_t for a given query/target utterance u_t based on its historical context $u_1, u_2, ..., u_{t-1}$ and the corresponding speaker information.

3.2 Overview

In this section, we present our multi-prompt and label paraphrasing (MPLP) model to mimic the thinking process. The overview of our model is shown in Fig. 1. Our model consists of two stages. The first stage is for utterance understanding, i.e., "1) *What does the speaker say*?". To this end, a PLM is fine-tuned to produce initial representations for utterances. The resulting model is saved as the base model (denoted as MPLP$_b$). The second stage is for the modeling of next three questions including "2) *What is the influence of the conversational context*?", "3) *What is the speaker's background*?", and "4) *How does the speaker feel*?". Accordingly, we construct a history-oriented prompt and an experience-oriented

[2] Our code and data are available at https://github.com/NLPWM-WHU/MPLP.


Dialogue History Sequence
h_0
h_1
...
h_{t-2}
h_{t-1}
w_0
w_1
w_1
w_0
h_t
$influ_t^{hist}$
h_t^{hist}
$\textbf{Prompt}_{hist}$ = E([s_t may feel [mask]]).
Training set
Retriever
Similar Sample Set
d_0
d_1
d_{k-1}
d_k
$influ_t^{exp}$
h_t^{exp}
$\textbf{Prompt}_{exp}$ = E([s_t may feel [mask]]).
G_t = SentiWordNet(Class).
BART Encoder
BART Decoder
BART Decoder
h'_t
$\textbf{Prompt}_{hist}$, $\textbf{Prompt}_{exp}$,
E([[SEP] s_t : What do you want me to do? … … * M: Calm yourself. *])
E([s_t feels [mask]]).
E([s_t feels G_t]) (G_t = frustrated, disappointedly unsuccessful.)

Class	Adjective	Paraphrase
0	Neutral	having no personal preference
1	Happy	enjoying or showing or marked by joy or pleasure
2	Angry	feeling or showing anger
3	**Frustrated**	**disappointingly unsuccessful**
4	Sad	bad; unfortunate
...	...	...

(b) History-oriented prompt

(c) Experience-oriented prompt

(d) Label paraphrasing

(a) An overview of our model at the second stage

Figure 1: An overview of our model. The first stage for utterance understanding is conventional and thus omitted, and we only present the structure used at the second stage (a). The utterance representations from the first stage are used to construct the history-oriented prompt (b) and the experience-oriented prompt (c). The label paraphrases from the SentiWordNet are used for the auxiliary generation task (d).

prompt based on the initial utterance representations from the first stage. Meanwhile, we perform an auxiliary task of label paraphrasing to leverage label semantics and fully elicit the lexical knowledge from the PLM.

3.3 Utterance Understanding

We adopt the generative PLM BART [Lewis *et al.*, 2020] for utterance understanding. We package the most recent m utterances and their corresponding speaker names along with the target sentence into a token sequence C_t, and feed it to the BART encoder. To distinguish the target utterance (s_t, u_t) from its context, a special token $*$ is added at the beginning and the end of the input utterance in the encoder. An emotional prompt P_t is sent to the decoder to get the representation of target utterance. E denotes the embedding layer:

$$C_t = [s_{t-m}, u_{t-m}, s_{t-m+1}, ..., *, s_t, u_t, *\,] \quad (1)$$

$$P_t = [s_t \text{ feels } \text{ [mask] }] \quad (2)$$

$$\mathbf{H}_t = \text{BART-Decoder}(E(P_t), \text{BART-Encoder}(E(C_t)) \quad (3)$$

The representation of the [mask] token in $\mathbf{H}_t$, denoted as $\mathbf{h}_t$, which reflects the underlying emotion understood by the model, will be used for training the model at the first stage with a cross entropy loss. After the first stage, we can obtain a preliminary understanding of the target utterance.

3.4 History-oriented Prompt Construction

Many studies [Hazarika *et al.*, 2018; Majumder *et al.*, 2019; Shen *et al.*, 2021b] have demonstrated the importance of historical information for the ERC task. However, the currently available prompt-based method CISPER [Yi *et al.*, 2022] simply constructs a shared prompt for all utterances in a dialogue. This hinders the model's ability to understand contextual information that is relevant to the target utterance. To address this issue, we propose a speaker-focused and history-oriented prompt generation method.

We generate a representation $\mathbf{h}_i$ for each historical utterance u_i using the fine-tuned BART obtained by the first stage. To concentrate on the utterances that are highly relevant to the target utterance, we calculate the representation similarity as the importance measure:

$$a_i^{hist} = \frac{exp(\mathbf{W}_h^{hist}[\mathbf{h}_i; \mathbf{h}_t])}{\sum_{i=0}^{t-1} exp(\mathbf{W}_h^{hist}[\mathbf{h}_i; \mathbf{h}_t])} \quad (4)$$

Following [Shen *et al.*, 2021b], a relation-aware feature transformation is applied to each historical utterance:

$$\mathbf{h}_i = \mathbf{W}_i^{hist}\mathbf{h}_i \quad (5)$$

where $\mathbf{W}_i^{hist} \in \{\mathbf{W}_0, \mathbf{W}_1\}$ is determined by whether a historical utterance u_i is of the same speaker with the target utterance u_t. This helps to distinguish the emotional effect of the current speaker and those of other speakers.

To make the emotional representation more contextualized, a Bi-LSTM module is applied to the historical sequence:

$$\tilde{\mathbf{h}}_i = \text{Bi-LSTM}(\tilde{\mathbf{h}}_{i-1}, \mathbf{h}_i) \quad (6)$$

Finally, we aggregate the historical emotional information to obtain the influence $\mathbf{influ}_t^{hist}$ of the conversational context on the current speaker. $\mathbf{influ}_t^{hist}$ is further added to the original utterance representation to capture the emotional impact of the dialogue history:

$$\mathbf{influ}_t^{hist} = \sum_{i=0}^{t-1} a_i^{hist}\tilde{\mathbf{h}}_i, \ \mathbf{h}_t^{hist} = \mathbf{influ}_t^{hist} + \mathbf{h_t} \quad (7)$$

Further, to allow the PLM to better utilize the target utterance-related history information, we construct a history-oriented prompt by replacing the embedding at the [mask] position in the original $E([s_t \text{ may feel [mask]}])$ prompt with $\mathbf{h}_t^{hist}$, and denote the resulting prompt as Prompt_{hist}:

$$\text{Prompt}_{hist} = E([s_t \text{ may feel}])\ \mathbf{h}_t^{hist} \tag{8}$$

Prompt_{hist} is appended to the input of the encoder at the second stage. During the training procedure, the history-oriented prompt is updated by dynamically selecting historical relevant information to continuously enhance the model.

3.5 Experience-oriented Prompt Construction

The speaker's background is also crucial to determine his/her attitude in the conversation. In particular, in multi-party conversations where the conversational context is less coherent, a speaker depends more on his/her experience in similar situations to facilitate conversation. In this section, we propose an experience-oriented prompt to encode the speaker's task-specific background.

We consider the training set, which has been seen by the model at the first stage as the speaker's task-specific background, and retrieve similar samples in it to build the experience-oriented prompt. To find similar samples, we use a text retriever such as BERTScore [Zhang *et al.*, 2020] and BM25 [Robertson and Zaragoza, 2009] to calculate the similarity between the target utterance u_t and an utterance u_d in the training set. The utterances with the top-k similarity are chosen as the similar sample set D.

$$sim(u_t, u_d) = \text{Retriever}(u_t, u_d) \tag{9}$$

By now, we have calculated the *text similarity* for the utterances themselves. However, as pointed out by previous work [Li *et al.*, 2022], even emotions of the same expression can vary dramatically in different context. To model the *similar situation* more precisely, we calculate the *context-influenced text similarity* between the similar samples and the target utterance, which is implemented by element-wise product of the utterance representations followed by a linear transformation. Since at the first stage, the local context is partially incorporated into the utterance encoding procedure, we consider their similarity as an indicator of the *context-influenced text similarity* between utterances:

$$a_j^{exp} = \frac{exp(\mathbf{W}_h^{exp}[\mathbf{d}_j \odot \mathbf{h}_t])}{\sum_{j=0}^{k} exp(\mathbf{W}_h^{exp}[\mathbf{d}_j \odot \mathbf{h}_t])} \tag{10}$$

$$\mathbf{influ}_t^{exp} = \sum_{j=0}^{k} a_j^{exp}\mathbf{d}_j,\ \mathbf{h}_t^{exp} = \mathbf{influ}_t^{exp} + \mathbf{h}_t \tag{11}$$

After the above two steps, we can use the training samples which are similar to the current utterance and have similar context as our *prior experience* to get a deep understanding of the speaker's background. Similar to the previous section, we use $\mathbf{h}_t^{exp}$ to construct the experience-oriented prompt:

$$\text{Prompt}_{exp} = E([s_t \text{ may feel}])\ \mathbf{h}_t^{exp} \tag{12}$$

Prompt_{exp} is also appended to the input of the encoder at the second stage to provide the speaker's background knowledge.

3.6 Label Paraphrasing

The rich semantics of labels are indispensable for distinguishing the subtle difference between labels. They are also beneficial to capture the text-label correlation. In view of this, we perform an auxiliary label paraphrasing task to assist the main emotion recognition task. We use the label names and their paraphrases in SentiWordNet 3.0 [3] to conduct the label paraphrasing task. To be specific, for a given label, such as *sadness*, we map it to the corresponding adjective and generates the sense gloss G_t, which is the gloss of the most frequent sense [4]. Finally, the input of the encoder at the second stage is denoted as C'_t:

$$\begin{aligned} I'_t = {} & \text{Prompt}_{hist},\ \text{Prompt}_{exp}, \\ & E([\text{[SEP]}\ s_{t-k}, u_{t-k}, ..., *, s_t, u_t, *]) \end{aligned} \tag{13}$$

There are two generative targets P_t and G_t on the decoder side. These two targets are fed into the decoder for two passes to get $\mathbf{H}'_t$ and $\mathbf{H}'_{gt}$, which are used for emotion classification and label paraphrase generation, respectively:

$$\mathbf{H}'_t = \text{BART-Decoder}(E(P_t), \text{BART-Encoder}(I'_t)) \tag{14}$$

$$\mathbf{H}'_{gt} = \text{BART-Decoder}(E([s_t \text{ feels } G_t]), \text{BART-Encoder}(I'_t)) \tag{15}$$

3.7 Training and Prediction

For the model training at the second stage, we take the representation of [mask] in $\mathbf{H}'_t$ as the final representation $\mathbf{h}'_t$, and apply a feed-forward neural network to get the predicted emotion logits p_t and the predicted label $\hat{y_t}$:

$$\mathbf{z}_t = \text{GeLU}(\mathbf{W}_H \mathbf{h'}_t + \mathbf{b}_H) \tag{16}$$

$$p_t = \text{Softmax}(\mathbf{W}_z \mathbf{z}_t + \mathbf{b}_z) \tag{17}$$

$$\hat{y_t} = \text{argmax}_{e \in Y}(p_t) \tag{18}$$

To fine-tune the model, the cross-entropy loss is used as the objective function:

$$L_{CE}(\theta) = -\sum_{j=1}^{M}\sum_{i=1}^{N_j} \log(p_{j,i}[y_{j,i}]) \tag{19}$$

where M is the number of conversations in the training set, N_j is the number of utterances in the j-th dialogue, and θ is the collection of trainable parameters in our model.

The auxiliary loss for generating label paraphrases is calculated and added via weighted sum:

$$L_{GEN}(\theta) = -\sum_{j=1}^{M}\sum_{i=1}^{N_j}\sum_{r=1}^{|G_{j,i}|} \log p(g_{r+1}|g_r, \theta) \tag{20}$$

$$L(\theta) = L_{CE}(\theta) + \alpha * L_{GEN}(\theta) \tag{21}$$

where g_r denotes the r-th token in the label gloss $G_{j,i}$, and α is a balancing weight for the loss of the label paraphrasing task.

[3] Note that our model does not need paraphrase for inference.

[4] We filter the paraphrases with both $PosScore$ and $NegScore$ equal to 0. Compared to the class category id and the adjective, the gloss contains more emotion-related words that are valuable for a deep label understanding.

4 Experimental Settings

4.1 Datasets and Metrics

We conduct experiments on three widely used ERC datasets, including MELD [Poria *et al.*, 2019a], IEMOCAP [Busso *et al.*, 2008], and DailyDialog [Li *et al.*, 2017].

MELD is collected from the TV show *Friends*. It consists of multi-party conversations and there are 7 emotion labels including *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear*.

IEMOCAP is a multimodal dyadic conversation dataset where each conversation is performed by two actors. There are 6 types of emotion, namely, $neutral$, $happiness$, $sadness$, $anger$, $frustrated$, and $excited$.

DailyDialog is a large collection of daily dialogues. In each conversation, there are two speakers. Each utterance is classified as *neutral*, *happiness*, *sadness*, *anger*, *surprise*, *disgust*, and *fear*. Over 83% of utterances in DailyDialog are classified as *neutral*.

Only the textual modal information is used in our experiments. We adopt the micro-averaged F1 excluding the majority neutral class for DailyDialog and the weighted-average F1 for other two datasets as metrics [Shen *et al.*, 2021b].

4.2 Baselines

We adopt 10 state-of-the-art baselines and divide them into two types: with or without external commonsense knowledge during inference. Our model belongs to the latter type.

DialogueRNN [Majumder *et al.*, 2019] uses three GRUs to keep track of the speaker states, proceeding contexts, and proceeding emotion.

DialogueGCN [Ghosal *et al.*, 2019] utilizes a graph-based structure to model self- and inter-speaker dependency of the interlocutors within a conversation.

DialogXL [Shen *et al.*, 2021a] modifies the XLNet [Yang *et al.*, 2019] with dialog-aware self-attention to capture useful intra- and inter-speaker dependencies.

DAG-ERC [Shen *et al.*, 2021b] exploits a directed acyclic graph to model the information flow from both long-distance and nearby context in a conversation.

CoG-BART [Li *et al.*, 2022] employs the BART-Large model, and augments it with supervised contrastive learning and response generation to facilitate dialogue understanding.

COSMIC [Ghosal *et al.*, 2020] is the first model that incorporates different elements of commonsense and leverages them to update conversation states.

TODKAT [Zhu *et al.*, 2021] combines topic information to help the model choose commonsense that is more relevant to the topic of current conversation.

SKAIG [Li *et al.*, 2021] builds a locally connected graph and classifies commonsense elements into present, past, and future types to enhance emotion transitions in the graph.

CauAIN [Zhao *et al.*, 2022] treats commonsense as the cause of emotion and adopts the attention mechanism to connect utterances.

CISPER [Yi *et al.*, 2022] is the first to leverage prompt learning for ERC. The context and commonsense of the entire conversation are encoded into shared prompts for utterances in a dialogue.

	Model	**MELD**	**IEMOCAP**	**DailyDialog**
without commonsense	DialogueRNN	63.61	64.76	57.32
	DialogueGCN	63.02	64.91	57.52
	DialogXL	62.41	65.94	54.93
	DAG-ERC*	63.37	**67.10**	58.25
	CoG-BART	64.81	66.18	56.29
with commonsense	COSMIC	65.21	65.28	58.48
	TODKAT	65.47	61.33	58.47
	SKAIG	65.18	66.96	59.75
	CauAIN*	65.15	64.29	57.08
	CISPER*	64.99	55.20	56.80
Ours	MPLP$_b$	65.46	64.85	57.06
	MPLP	**66.51**	66.65	**59.92**

Table 1: The main comparison results. For the models that use the test set for the checkpoint selection, we re-implement their official code and use the validation set for checkpoint selection for fair comparison. The results are marked with * in this table.

4.3 Implementation

The training process is divided into two stages. At the first stage, we fine-tune the BART-Larget model for a batch size of 8 utterances. Following [Li *et al.*, 2022], the AdamW optimizer is adopted with a learning rate of 2e-5 with a linear scheduled warm-up strategy. Our model is trained 4, 10, and 4 epochs for MELD, IEMOCAP, and DailyDialog, and the maximum input text length is set to 128, 160, and 128, respectively. After that, we start the prompting and paraphrasing for an additional 1 epoch at the second stage. The size of retrieved similar samples and the paraphrasing loss ratio is set via grid search. We use BERTScore as the text retriever for MELD and IEMOCAP, and BM25 for DailyDialog. The results on the test set come from the best checkpoint in the validation set. All experiments are performed on a single GeForce RTX 3090 GPU and are averaged over 3 runs.

5 Results and Analysis

5.1 Main Comparison Results

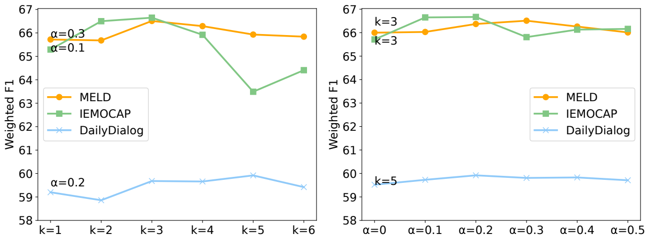
The main comparison results are shown in Table 1. In general, our MPLP model achieves the best performance on MELD and DailyDialog datasets, and its performance on IEMOCAP is also very competitive.

MELD The models based on commonsense knowledge perform better on this dataset. This suggests that commonsense knowledge can provide additional information and facilitate the understanding of the utterances. Moreover, our base model MPLP$_b$, which does not utilize commonsense knowledge, can achieve almost the same result as the best baseline TODKAT, indicating the strong utterance understanding capability of the PLM itself. Finally, with our proposed prompts and label paraphrasing, our complete model outperforms TODKAT by an absolute 1.04 F1 increase. This clearly demonstrates that our human-mimicking process can get a more comprehensive understanding of dialogue utterances.

IEMOCAP On this dataset, the graph-based models such as DAG-ERC and SKAIG can produce good results. The

	MELD	IEMOCAP	DailyDialog
Full Model	**66.51**	**66.65**	**59.92**
w/o Hist Prompt	65.76(↓0.75)	65.28(↓1.37)	58.60(↓1.32)
w/o Exp Prompt	65.77(↓0.74)	65.39(↓1.26)	59.38(↓0.54)
w/o Label Para.	66.00(↓0.51)	65.81(↓0.84)	59.52(↓0.40)

Table 2: The results for ablation study.



(a) (b)

Figure 2: The results for parameter study.

Model	MELD	IEMOCAP	DialyDialog
MPLP	**66.51**	**66.65**	**59.92**
rep. Add	65.67(↓0.84)	64.99(↓1.66)	58.12(↓1.80)
rep. Concatenate	65.00(↓1.51)	65.55(↓1.10)	57.11(↓2.81)

Table 3: The comparison results by using prompts and using features

main reason is that the conversations in IEMOCAP is extremely long [Shen *et al.*, 2021b], requiring complex stacked graph structures to model the dependencies of distant dialogues. Note that the performance of CISPER, which is also based on prompt, is extremely poor on this dataset. In contrast, our model closely follows behind DAG-ERC and SKAIG and is better than all other methods, indicating that our proposed prompt and paraphrasing learning mechanism can compensate for the issue of complicated dialogue structure modeling.

DailyDialog It is hard to recognize emotion from daily dialogues, and thus all methods produce worse results on this dataset than those on other two datasets. Nevertheless, our model outperforms the models without commonsense by at least an absolute 1.67 F1 increase. It also surpasses the models with commonsense and future information, showing the enormous understanding capacity of our proposed human-mimicking framework.

5.2 Ablation Analysis

To verify the effectiveness of each component in our model, we conduct a series of ablation study, by removing the history-oriented prompt (denoted as *w/o Hist Prompt*), the experience-oriented prompt (denoted as *w/o Exp Prompt*), and the label paraphrasing task (denoted as *w/o Label Para*) from the complete MPLP. The results are shown in Table 2. As we can observe, the performance drops on all datasets when each of the components is removed from the model. This proves the effectiveness of our proposed framework.

The history-oriented prompt yields the greatest impacts on three datasets. This is consistent with the findings in previous studies. Notably, the experience-oriented prompt makes almost the same contribution as the conversational context on MELD and IEMOCAP. This demonstrates that the modeling of speaker's background, which is first attempt made by us for the ERC task, is indeed essential for the understanding of the target utterance in most cases. The influence of experience-oriented prompt is not that significant on DailyDialog. The reason might be that daily conversations are often diversified, and it is hard to trace the speaker's background from the training set.

Label paraphrasing shows the greatest impact on IEMOCAP. This might be due to that the emotion labels in this dataset are much ambiguous. For example, it is hard to distinguish $sadness$, $anger$, and $frustrated$ since these types of emotion are often mixed together, and thus label paraphrasing helps a lot in disambiguation.

5.3 Parameter Study

There are two parameters in our framework: the number of similar samples k and the ratio of paraphrase generation loss α. In this section, we investigate the impact of these two parameters. The results are drawn in Fig. 2. We find that the trends on different datasets are similar with the change of parameters. They first rise to a peak and then fall gradually. A small k does not provide enough experience. When k is too large, the model is prone to introduce too much noise since more dissimilar sentences are added. The model with a small α can hardly learn label-semantics related information. If α is too large, the model may excessively emphasize the label semantics and ignores other factors.

5.4 Using Prompts *vs.* Using Features

We investigate the way in using the historical and background information, e.g., is prompt based learning a better way or can we simply use the same information as features?

To this end, we directly fuse the history or experience influenced representations to the utterance embeddings instead of using them as prompts. To be specific, we add (denoted as *rep. Add*) or concatenate (denoted as *rep. Concatenate*) these representations to the decoder output for final classification. In the first case, the classification head of the first stage is reused. In the second case, we retrain a new classification head with more epochs for a fair comparison since the feature dimensionality increases. The results are shown in Table 3.

As can be seen, though the same representations from conversational context and the speaker's background are fused into the final embeddings, both the performance of *rep. Add* and that of *rep. Concatenate* are consistently worse than our MPLP across all datasets. This indicates that the prompt based learning is a better way to inject relevant knowledge since the prompts are more understandable to a PLM than symbolic features.

5.5 Paraphrase Design

In this section, we examine the impact of paraphrase design on the model performance. We employ three experimental setups. The default one is our proposed MPLP model in this

Conversational Context	Similar Sample Set	DAG-ERC	CISPER	Ours
...... Chandler: I mean, that guy with the toe thing? (anger) Chandler: Who's he sleeping with? (neutral) **Chandler: Oh, c'mon Dora, don't be mad... (neutral)**	(1) Rachel: Mom, c'mon, stop worrying. (fear) (2) Rachel: Hey, c'mon, cut it out. (joy) (3) Rachel: Oh, c'mon. She's a person, you can do it! (joy)	sadness ✗	anger ✗	neutral ✓
Monica: I can't find garbage bags! (anger) Rachel: Oh, I think I saw some in here. (neutral) Monica: What is it?! (surprise)	(1) Chandler: I can't figure this out! (anger) (2) Rachel: I can't watch! (fear) (3) Monica: I can't do it! (sadness)	fear ✗	sadness ✗	anger ✓
...... Phoebe: Ohh, let me see it! Let me see your hand! (surprise) Monica: Why do you want to see my hand? (neutral) **Phoebe: I wanna see what's in your hand. I wanna see the trash. (disgust)** …...	(1) Charlie: I was (surprise) (2) Chandler: I like her. (neutral) (3) Woman: I love your car. (joy)	neutral ✗	neutral ✗	disgust ✓

Table 4: Case study on MELD. The target utterance in the conversational context is underlined.

Model	MELD	IEMOCAP	DailyDialog
MPLP	**66.51**	**66.65**	59.92
rep. Special Token	65.72(↓0.79)	65.01(↓1.64)	59.75(↓0.17)
rep. Label Adjective	66.21(↓0.30)	65.72(↓0.93)	**60.10**(↑0.18)

Table 5: The impact of paraphrase design.

paper, which uses the gloss of the target label in SentiWordNet as the generative target. The second one uses a special token for a label (denoted as *rep. Special Token*). The third one adopts the corresponding adjective of the emotion label (denoted as *rep. Label Adjective*) as the target. The results are shown in Table 5.

We find that the performance of *rep. Special Token* variant declines dramatically on all datasets. This is because the randomly initialized special tokens contain no semantics and cannot provide a proper guidance for the model. Compared to the special token, the performance of *rep. Label Adjective* variant is better, demonstrating the effectiveness of label understanding. The performance can be further improved by exploiting label glosses from SentiWordNet on MELD and IEMOCAP. However, the label's adjective works slightly better than the label paraphrasing on DailyDialog. This might be due to the diversity of the utterances in this dataset, where the the most frequent paraphrase of the emotion label cannot well satisfy such requirement and the corresponding adjective provides a more precise meaning for the label. Overall, label paraphrasing is helpful in most cases, yet its design can be further optimized, which we leave for the future work.

5.6 Case Study

To have a close look at the impact of the proposed prompts and label paraphrasing mechanism in our framework, We conduct a case study and show the results in Table 4.

The first case shows the effectiveness of history-oriented prompt. DAG-ERC misclassifies the utterance as $sadness$, showing that it fails to capture the context. CISPER incorporates the context but it pays more attention on the utterance with $anger$ emotion. In contrast, our model can select relevant information and make the correct prediction of $neutral$. To further confirm the effect of the history-oriented prompt, we replace our prompt with feature concatenation (Sec. 5.4), and find this variant also produces an $anger$ label.

The second case demonstrates the effectiveness of experience-oriented prompt. As the first utterance in the dialogue, the target utterance does not have any history information. As a result, both DAG-ERC and CISPER are unable to accurately recognize its emotion. In contrast, our model can make a correct prediction based on its prior experience. As can be seen, there is a utterance in the similar sample set which conveys the same emotion of $angry$ as that for the target utterance.

The third case emphasizes the importance of label paraphrasing, where no historical utterance or similar sample is related to the emotion of *disgust*. DAG-ERC and CISPER directly classify the utterance as $neutral$. In contrast, the word "trash", which is associated with the label's semantics of "disgust", helps our model to make a correct prediction.

In summary, the history-oriented prompt, the experience-oriented prompt, and the label paraphrasing improve our model's capability to recognize emotion in conversations.

6 Conclusion

We propose a novel framework for the ERC task which mimics the thinking process of a human being. We realize this process with a history-oriented prompt, an experience-oriented prompt, and the label paraphrasing mechanism, which can improve the understanding of the conversational context, the speaker's background, and the label semantics, respectively. We conduct experiments on three datasets, the results show that our method achieves competitive performance with the state-of-the-art baselines, proving the necessity of the modeling of human-thinking process, especially the understanding of the speaker's background, which has not been touched by existing studies. The ablation study and in-depth analysis further confirm the importance and effectiveness of using prompts and paraphrasing in our framework.

Ethical Statement

Emotion Recognition in Conversation has been a popular research topic in the academic community for a long time. The data and code used in our experiments are all open-source resources for research purposes.

Acknowledgements

This work was supported by the grant from the National Natural Science Foundation of China (NSFC) project (No. 62276193). It was also supported by the Joint Laboratory on Credit Science and Technology of CSCI-Wuhan University.

References

[Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 2010.

[Bao *et al.*, 2022] Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *IJCAI*, pages 4051–4057, 2022.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP*, pages 154–164, 2019.

[Ghosal *et al.*, 2020] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of EMNLP*, pages 2470–2481, 2020.

[Hazarika *et al.*, 2018] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP*, pages 2594–2604, 2018.

[Hu *et al.*, 2021] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *ACL/IJCNLP*, pages 7042–7052, 2021.

[König *et al.*, 2016] Alexandra König, Linda E. Francis, Aarti Malhotra, and Jesse Hoey. Defining affective identities in elderly nursing home residents for the design of an emotionally intelligent cognitive assistant. In *EAI*, pages 206–210, 2016.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880, 2020.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP*, pages 4582–4597, 2021.

[Li *et al.*, 2017] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pages 986–995, 2017.

[Li *et al.*, 2021] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of EMNLP*, pages 1204–1214, 2021.

[Li *et al.*, 2022] Shimin Li, Hang Yan, and Xipeng Qiu. Contrast and generation make BART a good dialogue emotion recognizer. In *AAAI*, pages 11002–11010, 2022.

[Liu *et al.*, 2021] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *CoRR*, abs/2103.10385, 2021.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN for emotion detection in conversations. In *AAAI*, pages 6818–6825, 2019.

[Majumder *et al.*, 2020] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979, 2020.

[Mueller *et al.*, 2022] Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. Label semantic aware pre-training for few-shot text classification. In *ACL*, pages 8318–8334, 2022.

[Poria *et al.*, 2019a] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pages 527–536, 2019.

[Poria *et al.*, 2019b] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[Pujol *et al.*, 2019] Francisco A. Pujol, Higinio Mora, and Ana Martínez. Emotion recognition to improve e-healthcare systems in smart cities. In Anna Visvizi and Miltiadis D. Lytras, editors, *RIIFORUM*, pages 245–254, 2019.

[Robertson and Zaragoza, 2009] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

[Sap *et al.*, 2019] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, pages 3027–3035, 2019.

[Shen *et al.*, 2021a] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *AAAI*, pages 13789–13797, 2021.

[Shen *et al.*, 2021b] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *ACL/IJCNLP*, pages 1551–1560, 2021.

[Song *et al.*, 2022] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. Supervised prototypical contrastive learning for emotion recognition in conversation. In *EMNLP*, pages 5197–5206, 2022.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764, 2019.

[Yang *et al.*, 2022] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. Hybrid curriculum learning for emotion recognition in conversation. In *AAAI*, pages 11595–11603, 2022.

[Yi *et al.*, 2022] Jingjie Yi, Deqing Yang, Siyu Yuan, Kaiyan Cao, Zhiyao Zhang, and Yanghua Xiao. Contextual information and commonsense based prompt for emotion recognition in conversation. In *ECML PKDD*, pages 707–723, 2022.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.

[Zhang *et al.*, 2021] Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. Aspect sentiment quad prediction as paraphrase generation. In *EMNLP*, pages 9209–9219, 2021.

[Zhao *et al.*, 2022] Weixiang Zhao, Yanyan Zhao, and Xin Lu. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*, pages 4524–4530, 2022.

[Zhu *et al.*, 2021] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *ACL/IJCNLP*, pages 1571–1582, 2021.